DOCUMENT RESUME

ED 435 702 TM 030 333

AUTHOR Fan, Xitao
TITLE Statistical Significance and Effect Size: Two Sides of a Coin.
PUB DATE 1999-11-00
NOTE 28p.; Paper presented at the Annual Meeting of the American Evaluation Association (Orlando, FL, November 3-6, 1999).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Decision Making; *Effect Size; Monte Carlo Methods; *Research Methodology; Sampling; Simulation; *Statistical Significance

ABSTRACT

This paper suggests that statistical significance testing and effect size are two sides of the same coin; they complement each other, but do not substitute for one another. Good research practice requires that both should be taken into consideration to make sound quantitative decisions. A Monte Carlo simulation experiment was conducted, and a three-factor crossed design, with 500 replications within each cell, was implemented in the simulation. The sampling variability of two popular effect sizes ("d" and "R squared") was empirically obtained under different data conditions. It is shown empirically that there is considerable variability of sample effect size measure, and the extent of sampling variability of effect size measures is strongly influenced by sample size. Although that which is statistically significant may not be practically meaningful, that which appears to be a practically meaningful effect size could occur by chance (i.e., sampling error), thus not trustworthy. It is pointed out that statistical significance testing and effect size measurement serve different purposes, and the sole reliance on either may be misleading. Some practical guidelines are recommended for combining statistical significance testing and effect size measure for making decisions in quantitative analysis. (Contains 2 tables, 3 figures, and 20 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 435 702

Effect_size v1.wpd

Statistical Significance and Effect Size: Two Sides of a Coin

Xitao Fan

Utah State University

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Xitao Fan

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Running Head: Significance and Effect Size

Send correspondence about this paper to:

Xitao Fan, Ph.D.
Department of Psychology
Utah State University
Logan, UT 84322-2810

Phone: (435)797-1451
Fax: (435)797-1448
E-Mail: fafan@cc.usu.edu

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM030333

Paper presented at 1999 American Evaluation Association Conference, November 3-6, Orlando, Florida (Session # 767).

BEST COPY AVAILABLE

Abstract

This paper argues that statistical significance testing and effect size are two related sides that together make a coin; they complement each other, but do not substitute one another. Good research practice requires that both should be taken into consideration in order to make sound quantitative decisions. A Monte Carlo simulation experiment was conducted, and a three-factor crossed design, with 500 replications within each cell, was implemented in the simulation. The sampling variability of two popular effect size measures (d and R^2) were empirically obtained under different data conditions.

It is shown empirically that there is considerable variability of sample effect size measure, and the extent of sampling variability of effect size measures is strongly influenced by sample size. Although what is statistically significant may not be practically meaningful, what appears to be a practically meaningful effect size could occur by chance (i.e., sampling error), thus not trustworthy. It is pointed out that statistical significance testing and effect size measure serve different purposes, and the sole reliance on either may be misleading. Some practical guidelines are recommended for combining statistical significance testing and effect size measure for making decisions in quantitative analysis.

In research and evaluation studies, statistical significance testing in quantitative research has received many valid criticisms in recent years, mainly for the reason that the outcome of statistical significance testing relies too heavily on sample size, and the issue of practical significance is often ignored. Consequently, such research practice limits understanding and applicability of quantitative research findings,. Effect size has been proposed as a supplement or alternative to statistical significance testing, and it has become increasingly popular. Some researchers, however, are not fully aware that, by itself, effect size may also be misleading, because sample size also has considerable influence on the sampling variability of effect size measures. This paper demonstrates through Monte Carlo simulation that statistical significance testing and effect size are two related sides that together make a coin; they complement each other, but not substitute to one another. Good research practice requires that both should be taken into consideration in order to make sound quantitative decisions. To lay a foundation for the discussion in this paper, some relevant issues related to statistical significance testing and effect size measures are first briefly reviewed.

Statistical Significance Testing

Use of Statistical Significance Testing in Research

There have been different misconceptions about what significance testing is, and what it is not (Shaver, 1993). For this paper, it is important to have a good understanding about the basic purpose of statistical significance testing in quantitative research, and about what information statistical significance testing provides for researchers.

The fundamental concept underlying statistical significance testing is sampling variation: from a population with known parameters (e.g., known population mean), sample statistics (e.g., observed sample mean) will vary around the population parameter to certain extent. How much

sampling variation can there be? How likely will an observed sample statistic (e.g., sample mean of 68) can occur due to sampling variability (i.e., "by chance") for a given population parameter (e.g., population mean of 80)? In a nutshell, statistical significance testing is conducted to evaluate the viability of null hypothesis by assessing how likely some observed sample statistic could have occurred as the result of random sampling variation for a given population parameter. More specifically, statistical significance testing answers the question: what is the probability of obtaining an observed sample statistic for a given or known population parameter?

Assuming that there exist two treatment conditions, A and B (e.g., A represents a new instructional approach in teaching mathematics, while B represents the conventional instructional approach currently in use). The program evaluation team is interested in knowing if A is better and more effective than B in teaching children math. The null hypothesis in this situation is that A are B are equal, i.e., students under A and B will learn equally well. Obviously, because of sampling variation, the two samples (one under A, and the other under B) typically will not have the same statistics, even if A is indeed the same as B. The question becomes: how different the sample statistics should be between A and B samples when we can say with confidence that A is different from B in effectiveness. Given the null hypothesis of no difference between A and B treatments, smaller observed difference between A and B samples is more likely to occur than larger observed difference between the two. When the difference between the two samples become sufficiently large relative to the theoretical random sampling variation such that it becomes highly unlikely if A and B are equally effective (null hypothesis of no difference), we conclude that the observed results is very unlikely to have occurred if the null hypothesis is indeed true. As a result, we reject the null hypothesis of no difference, conclude that A and B are not the same in their effectiveness in teaching math.

Note that in statistical significance testing, all we have assessed is the probability of obtaining the sample data (D) if the null hypothesis (H_0) is true, i.e., $p(D \mid H_0)$. If $p(D \mid H_0)$ is sufficiently small (e.g., smaller than .05 or .01), the null hypothesis will be considered not viable, and will be rejected. The rejection of the null hypothesis tells us that the random sampling variability is the unlikely explanation for the observed statistical results, but it gives no indication about how importance of our obtained statistical results. Going back to the example of A and B approaches in teaching mathematics, rejection of the null hypothesis (A and B are equally effective in teaching mathematics) simply tell us that it is unlikely that A and B are equally effective, but it does not give us any indication about how more effective A is than B, or vice versa. The real meaning of statistical significance testing, however, has often been lost in research practice, and the importance of statistical significance tends to be greatly exaggerated.

Major Criticisms of Statistical Significance Test

In research and evaluation studies, the over-reliance on statistical significance testing has been challenged on several grounds. Thompson (1993) discussed three relevant criticisms for statistical significance testing: over-dependency on sample size, some nonsensical comparisons, and some inescapable dilemmas created by statistical significance testing. In the similar vein, Kirk (1996) discussed three major criticisms of statistical significance testing: (1) significance testing does not tell researchers what they want to know, but rather, it creates the illusion of probabilistic proof by contradiction (Falk & Greenbaum, 1995); (2) statistical significance testing is often a trivial exercise, because it simply indicates the power of the design (which primarily depends on the sample size) to reject the false null hypothesis; and (3) significance testing "turns a continuum of uncertainty into a dichotomous reject-do-not-reject decision", and this dichotomous decision process may "lead to the anomalous situation in which two researchers obtain identical treatment

effects but draw different conclusions" (Kirk, 1996, p. 748) simply because of the slight differences in their design (i.e., sample sizes).

Of all the criticisms for statistical significance testing, probably the one best known is the over-reliance of statistical significance on sample size. It is well-known that the outcome of statistical significance testing heavily depends on the sample size used for the testing: for a fixed amount of difference between the hypothesized population parameter and the observed sample statistic, the larger the sample size, the easier it is to reject the null hypothesis. As discussed by Meehl (1978), ". . . the null hypothesis, taken literally, is always false" (p. 822). Because the null hypothesis is almost always theoretically false, statistical significance often becomes a matter of having sufficiently large sample in order to have enough statistical power for rejecting the null hypothesis. As Thompson (1992) sarcastically commented, in the ritualistic exercise of significance testing, ". . . tired researchers, having collected data from hundreds of subjects, then conduct a statistical test to evaluate whether there were a lot of subjects, which the researchers already know, because they collected the data and know they're tired." (p. 436).

Because the importance of statistical significance has traditionally been grossly exaggerated, statistical significance testing has become something sacredly ritualistic in quantitative analysis, to the point that statistical significance almost becomes the literal equivalent of importance of quantitative findings. Undoubtedly, this misconception has been compounded by the unfortunate misnomer of "significance" in this context.

Effect Size

The criticisms of statistical significance testing have led quantitative researchers to explore other approaches for making quantitative sense out of the data, because, as reasoned by many researchers (e.g., Kirk, 1996), the rejection of the null hypothesis by itself is not very informative.

There appears to be little doubt that the importance attributed to statistical significance testing in research and evaluation has traditionally far exceeded what is warranted in relation to the information such significance testing provides (e.g., Thompson, 1993).

Use of Effect Size Measure

Because statistical significance testing only shows in probabilistic terms how unlikely it is to obtain the sample data if the null hypothesis is true, but it does not inform whether the findings are practically meaningful or important, the general approach of obtaining some kind of scale-free effect size measure as the indicator of practical meaningfulness or importance has become popular, and its use in research practice has been widely advocated in recent years (Thompson, 1996). As the Publication Manual of the American Psychological Association (4th edition) explains, neither a priori nor exact probabilistic values reflect "the importance (magnitude) of an effect or the strength of a relationship because both probability values depend on sample size. You can estimate the magnitude of an effect with a number of measures that do not depend on sample size" (APA, 1994, p. 18).

Although there is some consensus that the role statistical significance testing plays in research practice should be reduced, and some other quantitative treatment of the data (e.g., effect size) should be used, there is less agreement about to what extent the role of statistical significance testing should be reduced, and to what extent the role of effect size should be enhanced in quantitative research. On one hand, statistical significance testing has been criticized as representing almost nothing but obstacles on our way of scientific inquiry (e.g., Carver, 1978, 1993; Meehl, 1978), as indicated by the strongly worded criticism that the reliance on significance testing for the null hypothesis "is a terrible mistake, a basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology" (Meehl, 1978, p.

817).

On the other hand, some other researchers has defended the legitimate role that the correct use of significance testing plays in scientific inquiry (e.g., Levin, 1993; Schafer, 1993). Levin (1993) argued that the baby (statistical significance testing) should not be thrown out with the bath water, just because the bath water might not be clean (misuse/misinterpretation of significance testing). Using hypothetical examples, Levin argued that, even with effect size measures, statistical significance testing is still essential in many situations so that researchers would not be misled by effect size measures.

Variety of Effect Size Measures

A variety of measures for effect size have been developed over the decades. Both Kirk (1996) and Snyder and Lawson (1993) provide useful and practical summary of the variety of measures of effect size. Because the terminology used for describing the variety of effect size measures has not been standard in the literature, sometimes there appears to be some confusion in reporting about what effect size measure has been reported in a study (Kirk, 1996). Maxwell and Delaney (1990) categorized the variety of measures of effect size into two broad categories: measures of effect size (based on group mean differences) and measures of the strength in association (based on proportion of variance accounted).

The first category, measures of effect size based on standardized group mean difference, is represented by Cohen's d, or some variations of it (e.g., Glass's g for meta-analysis and Hedges' g). In its most general sample form, d is expressed as: $d = \frac{\bar{X}_{group\ 1} - \bar{X}_{group\ 2}}{SD_{pooled}}$. Over the years, for research situations involving two groups where the comparison of the group means is the primary interest, d has become the measure of choice for effect size.

The second broad category, measures of the association strength that is based on the

proportion of variance accounted for, can be represented by $\underline{R}^2$ or η^2. The most general form for the association strength can be expressed as: $\eta^2 = \frac{Sum\ of\ Squares_{(a\ source)}}{Sum\ of\ Squares_{(Total)}}$. The numerator represents the sum of squares from a source of interest; and as such, either it may represent the sum of squares from one source out of multiple sources (e.g., sum of squares contributed by one predictor from a multiple-predictor regression model, or that contributed by one factor from a multi-factor analysis of variance model), or it may represent the sum of squares contributed by the full model. In the in the former case, $\underline{\eta}^2$ is typically used to quantify the proportion of variance accounted for by one factor (predictor), while in the latter case, $\underline{R}^2$ is usually used to quantify the proportion of variance accounted for by the full model. In this sense, $\underline{\eta}^2$ and $\underline{R}^2$ are basically the same thing.

Because $\underline{R}^2$ contains upward bias due to the maximization property of least square principle, different bias-corrected counterparts of $\underline{R}^2$ have been proposed, such as ω^2, ϵ^2, and others (see computational details in Kirk, 1996, and Snyder & Lawson, 1993). A literature review of several influential journals in psychology has shown that $\underline{R}^2$ is the most popular measure reported for measuring association strength (probably because of its availability from statistical software programs), while bias-corrected counterparts of $\underline{R}^2$ (e.g., ω^2, ϵ^2, and others) have been minimally reported (Kirk, 1996).

Effect Size as a Random Variable

In using effect size measure in research, an important dimension of any measure of effect size seem to have been ignored by many: effect size measure itself is a random variable, just as, say, sample mean is a random variable. Being a random variable has one important implication for its interpretation: if we are dealing with samples, effect size measure obtained from a sample is subject to sampling variability as dictated by its underlying sampling distribution. Furthermore,

the extent of sampling variability of an effect size measure is influenced by sample size from which the effect size is obtained, similar to the situation where the probability value of a test statistic in statistical significance testing is influenced by the sample size used to obtain the test statistic. In other words, when sample size is small, the sample effect size may deviate substantially more from the population effect size than when sample size is large.

Although the random variable nature of effect size measures has been widely known in the literature of quantitative analysis (e.g., Fowler, 1985; Hedges & Olkin, 1985; Glass & Hopkins, 1996, Chapter 14), relatively few practitioners pay any attention to, or show any interest in, this fact. In the research literature, it is not uncommon to encounter discussion to the effect that the outcome of a statistical significance test is heavily influenced by sample size (true!), so attention should be paid to effect size, as if effect size were not influenced by sample size. Undoubtedly, the use of effect size measure makes good quantitative and common sense, but quantitative practitioners should realize that the use of effect size serves a different purpose from that of statistical significance test: while statistical test evaluates the probability of obtaining the sample outcome by chance (sampling error), effect size provides some indication for practical meaningfulness. Although a statistically significant outcome may not be practically meaningful, what appears to be a practically meaningful outcome may also have occurred by chance, and consequently, it is not trustworthy.

The general purpose of this paper is to demonstrate that statistical significance testing and effect size measure are both needed for making sound quantitative decisions; they supplement each, but do not substitute each other, because the two serve different purposes. To accomplish the general purpose of the paper, the following specific objectives are to be addressed in the paper:

1. to empirically assess the extent of sampling variability of major effect size measures;
2. to empirically assess the influence of sample size on the extent of sampling variability of major effect size measures;
3. to offer some practical guidelines for using both statistical significance testing outcome (i.e., test statistic and the associated probability value) and the descriptive effect size measure for arriving at sound quantitative decisions in research and evaluation.

Methods

Although theoretical sampling distributions of some popular measures of effect size have been known [e.g., see Hedges & Olkin (1985) for d, Glass & Hopkins (1996) for R^2), empirical approach was adopted in this paper for the purpose of attaining better understanding of the variability of effect size measures under different data conditions. Monte Carlo method was used to simulate different data conditions under which both effect size measures and statistical significance testing outcomes were accumulated and later analyzed.

Design

Of the variety of effect size measures, in this paper, two most widely known effect size measures were used: d (standardized mean difference) and R^2 (proportion of variance accounted for). These two effect size measures are so well-known that it is probably safe to bet that those who have been exposed to the concept of effect size have seen one or both of these measures. The literature review of several psychology journals by Kirk (1996) indicates that R^2 is by far the most frequently reported effect size measure, probably because it is routinely reported in regression or general linear model procedures of statistical software. The meta-analysis work by Glass (1976) and its later wide spread use undoubtedly contributed to the popularity of the effect size measure d.

For evaluating d (standardized two-group mean difference), samples from two statistical populations with known population parameters were generated. Three factors were considered in the Monte Carlo simulation design: (a) 4 levels of population effect size (d = .00, .20, .50, and .80 respectively) that correspond to zero, small, medium, and large effects based on the guidelines suggested by Cohen (1988, p. Chapter 2); (b) 5 levels of sample size conditions (N=20, 40, 80, 160, 240); and (c) 4 conditions of group variability (as represented by standard deviation) ratio (σ_1/σ_2 = 1, 1.5, 2, 2.5, respectively). For the fully crossed design, these three factors yielded 80 (4×5×4) cells. Five hundred replications were conducted within each cell, making the total number of samples generated for evaluating d to be 40,000 (500×80).

Regression models were used for evaluating R^2 (proportion of variance accounted for). Three factors were considered in the design: (a) 4 levels of population effect size (R^2 = .00, .02, .12, and .25, respectively) that approximately correspond to zero, small, medium, and large effects according to the guidelines suggested by Cohen (1988, Chapter 9); (b) 4 levels of sample size conditions (N = 20, 40, 80, 160, respectively); and (c) 2 conditions for the number of predictors (k = 2 and 4, respectively), with the collinearity among the predictors set at .10. The fully crossed design of these three factors called for 32 cells (4×4×2). With 500 replications within each cell, the design required the generation of a total of 16,000 (32×500) samples. The designs for evaluating d and R^2 were graphically presented in Figure 1.

[Insert Figure 1 about here]

Data

Data generation was accomplished by using the SAS normal data generator. Multivariate normal data for regression models were simulated using the matrix decomposition procedure

(Kaiser & Dickman, 1962). All sample data generation, sample effect size calculation, and statistical significance testing were accomplished through the Interactive Matrix Language (PROC IML) of the SAS system (SAS Window Version 7.0). It should be noted that data non-normality was not considered in the present paper. As a result, the influence of data non-normality on both effect size measures and on statistical significance test outcomes was not assessed.

Results and Discussions

Figure 2 graphically presents the sampling variability (in the form of 90% confidence interval) of the effect size measure of standardized mean difference d for four conditions of population effects: zero, small, medium, and large (population d = .00, .20, .50, and .80, respectively). In addition to sample size conditions, the four conditions of group variability ratio(σ_1/σ_2) were also presented (σ_1/σ_2 =1, 1.5, 2, and 2.5). In Figure 1, a hi-low bar represents a 90% confidence interval of sample d for a condition of sample size and that of a group variability ratio (SD ratio: ratio of the standard deviations of two groups), and a short horizontal dash line within a bar represents the mean of 500 sample ds.

[Insert Figure 2 about here]

Several observations can be made from Figure 2. First, sample effect size measure d appears to be an unbiased estimate of population d. The characteristic of unbiasedness of sample d is obvious because the mean of sample d is very close to the known population value (.00, .20, .50, and .80, respectively) under most data conditions. However, larger discrepancy between the two population standard deviations of the two groups (SD ratio) causes some minor degree of downward bias of sample d, and this is especially obvious under the condition of population d =

.80.

Second, there is considerable sampling variability of sample d. For example, under the condition of population d = .00 (i.e., two samples drawn from the same population, thus no real difference between the two samples), for small size condition such as N=20 (n_1=n_2=10), the 90% confidence interval almost covers the range from -.80 to +.80. In other words, for this sample size condition, when two samples are drawn from the same population (i.e., there is absolutely no real difference between the two groups), we could have obtained what is typically considered as large effect size (±.80) just by chance (due to sampling error). Even when sample size is increased to N=80 (n_1=n_2=40), probably a moderate sample size condition for many experimental designs, we could still have obtained sample effect size almost as large as ±.40 (moderate effect) by chance.

Third, the extent of sampling variability is obviously affected by sample size. It is seen that, with the increase of sample size, the sampling variability of sample d, as represented by the 90% confidence intervals, shows a clear trend of becoming gradually smaller under all the conditions of population effect size (zero, small, medium, and large). This indicates that, if we have two identical effect sizes (e.g., moderate effect of d=.40) from two different studies involving different sample sizes [e.g., one is based on sample size of 40 (n_1=n_2=20), and the other is based on N=160 (n_1=n_2=80)], the one based on larger sample size is more trustworthy, because it is much less likely to have occurred due to sampling error or chance. This indicates that effect size measure should not be used by itself; instead, it should be considered together with sample size.

Figure 3 presents the sampling variability of another major type of effect size measure: measure of association strength as represented by the R^2. Because sample R^2 is widely known to

have upward bias, one form of bias-corrected $\underline{R}^2$ (adjusted $\underline{R}^2$ based on Wherry formula[1] that has been implemented in both SAS and SPSS regression procedure). The sampling variability of the $\underline{R}^2$ and adjusted $\underline{R}^2$ is represented by the 90% confidence interval bar, and the mean $\underline{R}^2$ based on 500 samples is represented by the dash line within each confidence interval bar.

[Insert Figure 3 about here]

In addition to some common observations already discussed for the effect size measure $\underline{d}$ in Figure 2, several observations unique for sample $\underline{R}^2$ in Figure 3 can be made. First, while sample $\underline{d}$ in Figure 2 has shown to be an unbiased estimator of population $\underline{d}$, sample $\underline{R}^2$ has obvious upward bias, as indicated by the position of mean $\underline{R}^2$ (dash line within each 90% confidence interval) that is consistently, and sometimes considerably, above the population $\underline{R}^2$ under all four population effect size magnitude conditions ($\underline{R}^2$=.00, .02, .12, and .25). Bias correction, however, has worked well for adjusted $\underline{R}^2$, with all means of sample adjusted $\underline{R}^2$s being very close to the population value.

Second, it is noted that $\underline{R}^2$ from the 4-predictor regression model has more upward bias than that from 2-predictor regression model. This finding is expected, because under the same sample size condition, the ratio of sample size to the number of predictors (N/p) is smaller for 4-predictor model than that for 2 predictor model. As is widely known, In regression analysis, it is often this ratio, rather than sample size per se, that largely determines the stability of regression analysis outcomes (Stevens, 1996).

---

[1] Adjusted $\underline{R}^2$ is obtained by: $\hat{R}^2 = 1 - \frac{N-1}{N-P-1}(1 - R^2)$. $\underline{R}^2$: uncorrected sample $\underline{R}^2$; N: sample size; P: # of predictors in the regression model.

Both $\underline{R}^2$ and adjusted $\underline{R}^2$ show considerable sampling variability, and the sampling variability of both decreases with the increase of sample size. The considerable sampling variability may make it relatively easy to obtain a medium and even large effect size measure by chance, even when the population effect size is zero or very small ($\underline{R}^2$=.02). For example, for population $\underline{R}^2$=.02 (very small effect), and for the 4-predictor regression model, the upper 90% confidence limit of sample $\underline{R}^2$ reaches as high as over .46 for N=20 (very large sample effect), and about .25 for N=40 (large effect). This degree of sampling variability that is strongly affected by sample size (or N/p ratio) highlights the need that effect size should be considered within the context of sample size; used by itself, sample effect size measure may be misleading.

Table 1 presents the percentages of statistically significant tests under different population effect size conditions, and under different sample size conditions. When the population effect is zero, approximately 5% of tests are statistically significant (underlined entries in the table), close to the specified nominal Type I error rate (α level). While population effect is not zero, the table entries represent the power of the statistical tests involved. It is seen that the tests can adequately detect [defined as statistical power about.80 (Stevens, 1996)] the population effect only when the population effect is moderate and large ($\underline{d}$=.50, .80), and the sample size is not small ($N \geq 40$) (see shaded entries in the table).

[Insert Table 1 about here]

While we do not want to trust something which may have occurred by chance (Type I error), Table 1 shows that statistical tests may cause concern of Type II error when population effect is non-zero, i.e., we conclude that there is no population effect when in fact there is. To

balance these two opposite logical errors requires that researcher understand the consequences of Type I and II error respectively, take into consideration of effect size measure, and make decisions accordingly. Table 2 offers some practical guidelines for combining statistical significance testing and effect size measure in quantitative analysis. The content of Table 2 is self-explanatory, thus requiring no explanation or discussion here.

[Insert Table 2 about here]

Summary and Conclusions

This paper attempts to demonstrate through Monte Carlo simulation that statistical significance testing and effect size are two related sides that together make a coin; they complement each other, but do not substitute one another. Good research practice requires that both should be taken into consideration in order to make sound quantitative decisions. The Monte Carlo simulation used three-factor crossed design, with 500 replications within each cell. The sampling variability of two popular effect size measures (d and R^2) were obtained through simulation under different data conditions (e.g., population effect size, sample size).

It is shown empirically that there is considerable variability of sample effect size measure, and the extent of sampling variability of effect size measures is strongly influenced by sample size. Due to the sampling variability of effect size, what appears to be practically meaningful effect size may be the result of sampling error, and consequently, it is not trustworthy. It is pointed out that statistical significance testing and effect size measure serve different purposes, and the sole reliance on either may be misleading . Some practical guidelines are recommended for combining statistical significance testing and effect size measure for making decisions in quantitative analysis.

References

American Psychological Association. (1994). Publication Manual of the American Psychological Association (4th edition). Washington DC: American Psychological Association.

Carver, R. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). New York: Lawrence Erlbaum.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. Theory & Psychology, 5, 75-98.

Fowler, R.J. (1985). Point estimate and confidence intervals in measures of association. Psychological Bulletin, 98, 160-165.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5, 3-8.

Glass, G. V., & Hopkins, K. D. (1996). Statistical methods in education and psychology, (3rd ed.). Boston, MA: Allen and Bacon.

Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis. Orlando, FL: Academic Press.

Kaiser, H. F., & Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. Psychometrika, 27, 179-182.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56, 746-759.

Levin, J. R. (1993). Statistical significance testing from three perspectives. Journal of Experimental Education, 61, 378-382.

Maxwell, S. E., & Delaney, H. D. (1990). Designing experiments and analyzing data: A model comparison perspective. Belmont, CA: Wadsworth.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

Schafer, W. D. (1993). Interpreting statistical significance and nonsignificance. Journal of Experimental Education, 61, 383-387.

Shaver, J. P. (1993). What statistical significance testing is, and what it is not. Journal of Experimental Education, 61, 293-316.

Snyder, P., & Lawson, S. (1993). Effect size estimates. Journal of Experimental Education, 61, 334-349.

Stevens, J. (1996). Applied multivariate statistics for the social sciences. Mahwah, NJ: Lawrence Erlbaum Associates.

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. Journal of Counseling and Development, 70, 434-438.

Thompson, B. (1993). The use of statistical significance in research: Bootstrap and other alternatives. Journal of Experimental Education, 61, 361-377.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25, 26-30.

Table 1: Percentages of Statistically Significant Tests (α=.05)

	Population D			
Sample N ($N=n_1+n_2$; $n_1=n_2$)	.00	.20	.50	.80
20	5.90	7.35	18.05	37.25
40	5.90	8.85	32.75	65.30
80	5.25	14.00	54.40	92.40
160	5.95	22.50	85.30	99.75
240	5.65	32.80	96.45	99.95

	Population R^2			
Sample N	.00	.02	.12	.25
20	4.10	7.60	21.00	47.30
40	6.50	9.50	44.40	82.00
80	5.70	18.70	77.10	98.60
160	4.00	31.20	98.10	100.00

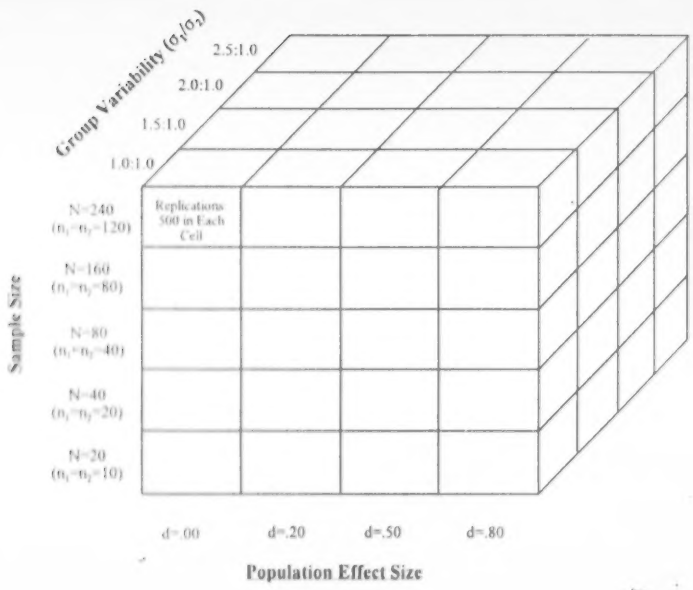
BEST COPY AVAILABLE

Table 2 Guidelines for Combining Significance Test Outcome with Effect Size Measure

Statistical Significance	Effect Size: Small	Effect Size: Medium	Effect Size: Large
No	1. It appears that there is no statistical nor practical effect. 2. Unless future research indicates otherwise, null hypothesis is favored in both statistical and practical sense.	1. If Type I error is your major concern, caution is warranted in interpreting the effect size by itself, because medium effect size could have been the result of chance, even if it may look practically meaningful. 2. If Type II error is your major concern, the sample effect looks promising, and you may take a closer look at the power of your test, because if your sample size is small, you may not have the statistical power to detect potential meaningful effect.	1. If Type I error is the major concern, some caution is still needed, because large effect size could have occurred by chance when sample size is small. Despite the needed caution , you have some evidence that meaningful population effect may exist. 2. If Type II error is your major concern, you may take a critical look at the power (more likely, lack thereof) of your statistical test. 3. Tentatively favor the practical significance of the effect, while keeping an open mind for further research findings .
Yes	1. The statistical significance is not accompanied by practical significance, and could have been the result of statistical power (large N). 2. Considerable caution is warranted, and statistical significance should not be interpreted as meaning something practically meaningful.	1. It is very unlikely that the observed effect is due to statistical chance. 2. The magnitude of effect is practically meaningful in many areas of social and behavioral sciences. 3. Conclude that effect is meaningful both statistically and practically.	1. There is high degree of certainty that the observed effect is not due to chance statistically, and the magnitude of the effect is also practically meaningful. 2. Conclude with confidence that effect is meaningful in both statistical and practical sense.

Note: The cell shade indicates degree of certainty about statistical and practical meaningfulness of the observed effect, with the darker shade indicates higher degree of certainty.

BEST COPY AVAILABLE



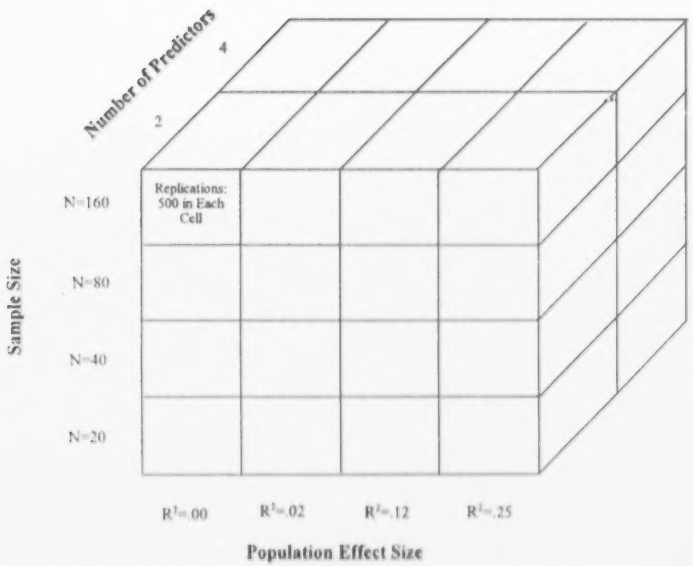


Figure 1: Study Design for Effect Size Measures of d and R^2

BEST COPY AVAILABLE

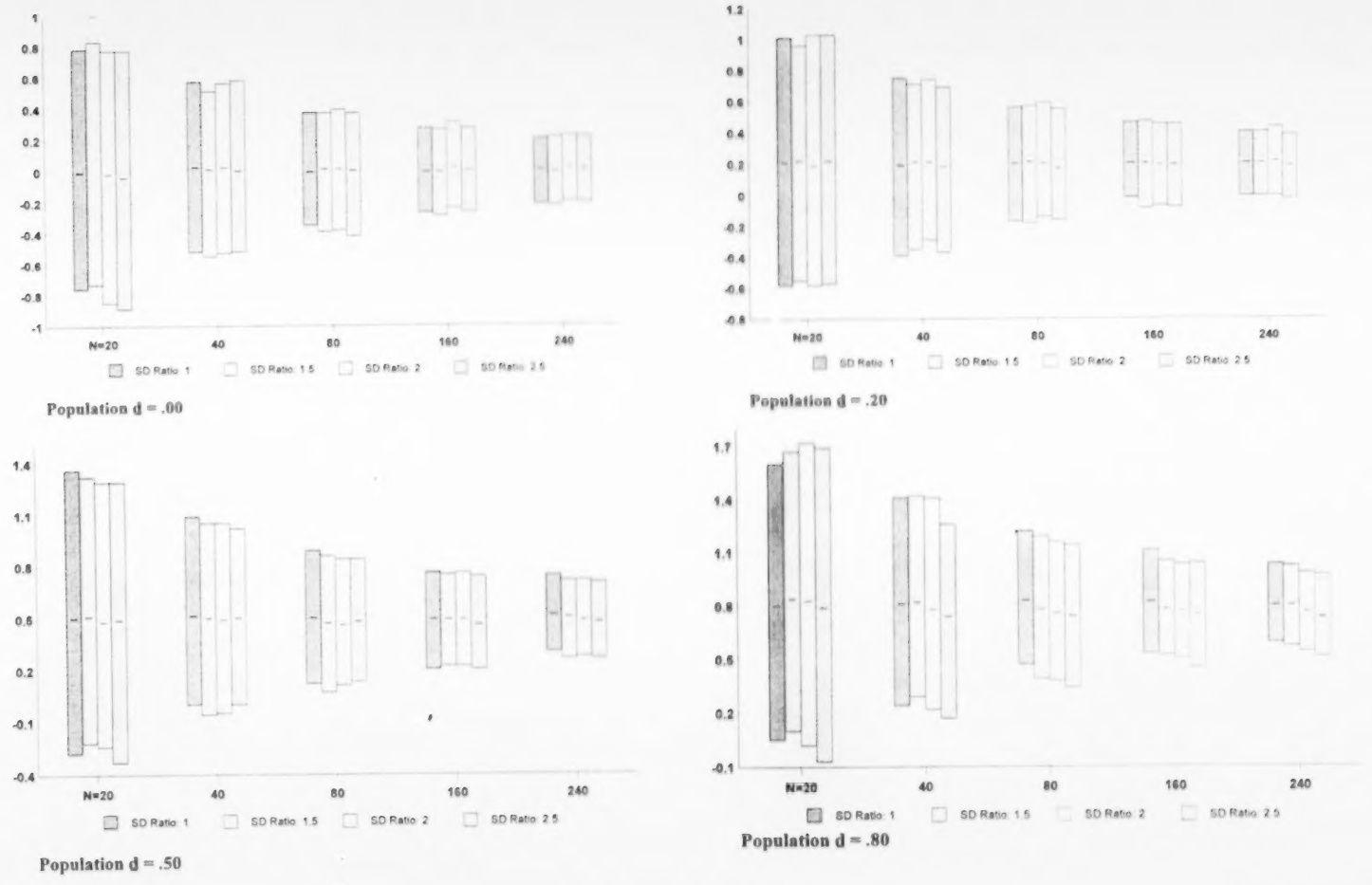


Figure 2: Confidence Intervals (90%) of Sample Effect Size d for Four Conditions of Population Effects

BEST COPY AVAILABLE

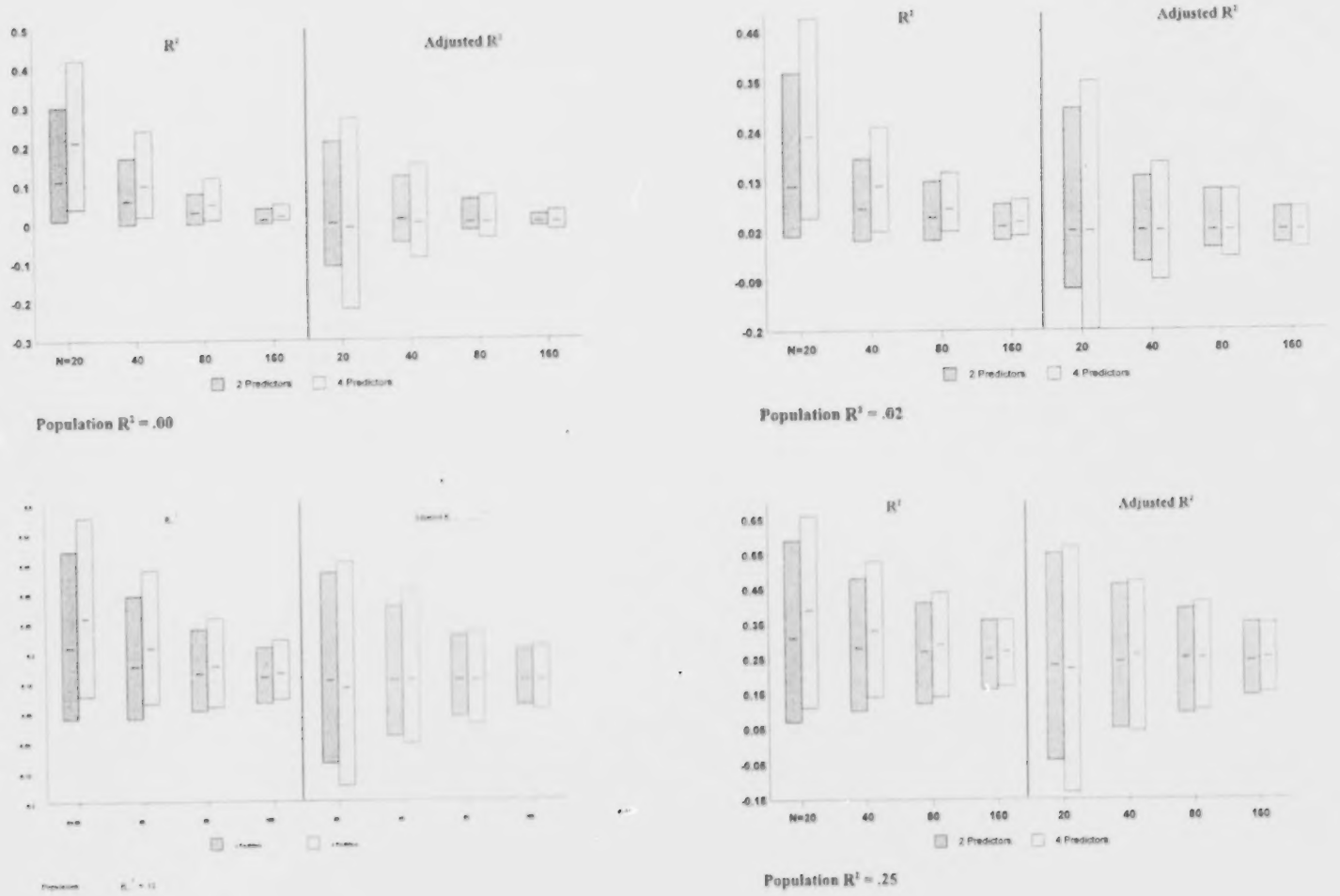


Figure 3: Confidence Intervals (90%) of Sample Effect Size $\underline{R}^2$ for Four Conditions of Population Effects

BEST COPY AVAILABLE